View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PUblication MAnagement

Deliverable D4.2
Augmented version of the bio-lexicon extended with bio event information and term-to-term weighted links (report + bio-lexicon)
Final version

Project acronym:	BOOTStrep
Project full title:	Bootstrapping Of Ontologies and Terminologies STrategic REsearch Project
Proposal/Contract no.:	FP6 - 028099
Duration:	April 01, 2006 – March 31, 2009
Project coordinator:	FSU Jena
Website:	www.bootstrep.eu
Authors:	Simonetta Montemagni, Giulia Venturi (CNR-ILC); Paul Thompson, Yutaka Sasaki, Sophia Ananiadou, John McNaught (UoM); Jung-jae Kim, Dietrich Rebholz (EBI)
Date of preparation:	30 September 2008
Nature	R+P
Dissemination level:	PU

Table of Contents

Summary ..... - 1 -
1 Introduction ..... - 2 -
2 Corpus annotation ..... - 3 -
2.1 Bio-event linguistic annotation ..... - 3 -
2.1.1 Corpus revision ..... - 3 -
2.2 Final Annotation results ..... - 14 -
2.2.1 Corpus Statistics ..... - 15 -
2.3 Modality annotation ..... - 25 -
2.3.1 Testing the classification scheme ..... - 25 -
2.3.2 Results ..... - 27 -
2.3.2.1 *Knowledge Type* information ..... - 30 -
2.3.2.2 *Certainty* information ..... - 31 -
2.3.2.3 *Point Of View* information ..... - 34 -
2.3.3 Summary of modality results ..... - 36 -
3 Extraction of bio-event frames from the BELA corpus ..... - 37 -
3.1 Event Patterns ..... - 37 -
3.2 Event frames ..... - 38 -
3.3 Corpus Format ..... - 39 -
3.4 Event Frame Extraction ..... - 40 -
3.5 NER ..... - 41 -
3.6 Semantic Role Labeling ..... - 41 -
3.7 Event pattern matching ..... - 42 -
3.8 Experimental Results ..... - 43 -
4 Representation of acquired bio-event frames in the Bio-Lexicon ..... - 44 -
5 Syntax-semantics linking ..... - 47 -
5.1 The starting point ..... - 47 -
5.2 Approaching the problem ..... - 48 -
5.2.1 Analysis of the syntax-semantics linking literature ..... - 50 -
5.2.2 A list of ‘prototypic’ syntactic realisations of semantic arguments ..... - 52 -
5.2.3 General language repositories of semantic frames ..... - 54 -
5.2.4 Annotated corpus ..... - 55 -
5.3 The mapping results ..... - 55 -
5.3.1 Full mapping ..... - 57 -
5.3.2 Partial mapping ..... - 58 -
5.3.3 Other results ..... - 60 -
5.4 Representation of syntax-semantics linking in BL ..... - 60 -
6 Linking between BELA and BEBA corpora ..... - 63 -
7 References ..... - 66 -

Summary

The main goal of WP4 was the semi-automatic acquisition of information about “bio-events” from biomedical literature. In this report, the final outcome of the work package is described, i.e. the version of the Bio-lexicon populated with verbs and nouns (nominalised verbs) expressing the most salient event relations between terms. This report is intended to complement and document the information contained in the augmented version of the Bio-Lexicon delivered as part of D4.2.

1 Introduction

The main goal of WP4 was the semi-automatic acquisition of information about "bio-events" from biomedical literature. In this report the final outcome of the work package will be described, i.e. the version of the Bio-lexicon populated with verbs and nouns (nominalised verbs) expressing the most salient event relations between terms. This report is intended to complement and document the information contained in the augmented version of the Bio-Lexicon delivered as part of D4.2.

A remark is in order here to clarify the misalignment between the title of the deliverable and its content. According to the title, the augmented version of the bio-lexicon was expected to be extended at this stage with bio-event information as well as with term-to-term weighted links. This is what was foreseen in the Technical Annex where it was stated that "identified bio-events will also be used to compute term-to-term similarity scores, based on the events in which they participate. These similarity scores between terms will be stored in the Bio-Lexicon and will represent useful input for the lexicon-to-ontology mapping (WP05) and, in particular, for ontology tuning and extension." Actually, after first experiments in this direction in which term similarity was computed starting from corpus evidence, it turned out that the bootstrapped terminological clusters, in spite of their being linguistically sound, did not appear to convey useful information from the biological point of view. On the other hand, we identified a gap in the original TA of the project, which did not foresee any mapping between syntactic subcategorization frames acquired in the framework of WP3 and event frames extracted in the framework of WP4. It was thus decided to tackle the issue of the syntax-semantics mapping by exploring the possibility of semi-automatically linking extracted subcategorization frames with Bio-event frames. The results of this activity are reported in Section 5.

2 Corpus annotation

When D4.1 was delivered, annotation activities were still ongoing. This section summarises the results of the annotation work carried out in the framework of WP4. It includes information concerning the following:

- bio-event linguistic annotation and modality annotation (UoM and ILC);
- annotation of biological events on full texts (so-called human test corpus) (EBI).

2.1 *Bio-event linguistic annotation*

This section describes the final results of bio-event linguistic event annotation. Firstly, the process of corpus revision is described. This was carried out when the corpus collection had finished, in order to enhance the consistency of the annotations according to the guidelines, and to improve the syntactic correctness of the corpus. Secondly, we provide some statistics about events in the final version of the corpus, together with a detailed examination of inter-annotator agreement and consistency issues.

2.1.1 Corpus revision

After corpus annotation has finished, a phase of corpus revision was begun. Examination of the corpus revealed a number of errors, which we felt could be corrected in a fairly straightforward way. This correction phase would enhance the consistency of the corpus, with the aim of producing more accurate results in the automatic acquisition of event frames, as well as enhancing the potential value of the corpus for reuse in other tasks.

We have only carried out revisions that can be done fairly mechanically, and involve as little bias as possible. Thus, we have not considered issues such as whether a particular semantic argument should or should not have been included within an event, or whether the correct named entity category has been assigned to a span. These are subjective issues and making decisions on these ourselves would introduce our own bias into the corrected corpus. Instead, we concentrated on the following tasks:

1) Ensuring, as far as possible, that the annotations comply with the annotations guidelines
2) Correcting errors in annotator-added chunks.

Task 1) was carried out manually, and the main focus was on correcting the *forms* (i.e. spans) of the annotations. Task 2), on the other hand, was carried out largely automatically, using custom scripts to correct regular chunking errors, together with a small amount of manual revision.

Only if the form of the semantic argument is not as expected, according to the guidelines, do we consider the context in which it occurs within the relevant abstract. This allows us to verify whether the argument is indeed correct, or whether changes need to be made.

2.1.1.1 *Ensuring annotation consistency*

For the manual correction phase of ensuring that annotations are consistent with the guidelines, we took steps to prevent biased changes being made to the corpus, in that events were viewed out of the context of the abstracts from which they were derived. To facilitate this, a script was run which generated a text file containing just the slots and fillers of all events in the annotated corpus. A further advantage

to viewing events in this way is that it facilitates quicker review of events than if each abstract was opened individually in WordFreak; this was an important consideration given that the review was being carried out manually. An example of the format used for the events within the text file is shown below:

```
Event: becoming
Verb: becoming
Theme: The cspA mRNA level
Condition: mid-to-late exponential growth phase
Descriptive-Theme: virtually undetectable
```

For each event, the text span over which the event was created is labelled as “Event”. On the next line, the value of the “Verb” slot is shown, which should normally be the head verb in the case that the event was created over a verb phrase, or a nominalised verb in the case the event was created over a noun phrase. On subsequent lines, the semantic roles that were filled during the annotation of the event are displayed. If a semantic argument consists of multiple discontinuous spans, this is indicated by the presence of the string “[AND]” between the different parts of the span, e.g.:

```
Theme: deoxyribonucleic acid [AND] polymerase II
```

This format allows us to concentrate only on the forms of the phrases that constitute the semantic arguments, and the semantic roles assigned to them. Most of the errors we are looking for are concerned with potential errors in the form of the phrases that constitute the semantic role (e.g. a preposition being present at the beginning of the argument when the guidelines state that arguments belonging to most roles should not begin in this way). To a lesser extent, we are also looking for incorrectly assigned semantic roles. However, only a limited number of semantic roles are changed, and only if the form of the semantic argument makes it clear that it should be assigned to a different role type, according to the annotation guidelines. Further explanation of these cases is provided below.

If a potential error is noticed, the appropriate abstract is opened in WordFreak, and the text span from which the event was derived is viewed. This allows us to verify whether there is indeed a problem with the argument, and allows changes to be made to the annotation if necessary. It is important to note that viewing events in context (i.e. within the abstracts from which they are derived) is normally used only to make decisions about whether the extent of the text span that constitutes the semantic argument needs to be changes. Decisions about changing semantic roles are not made from this contextual view, as this again is likely to introduce some bias into the revised corpus.

2.1.1.2 *Changes made to the annotated corpus*

Verb-related changes

We made two main passes through the annotated events in order to identify and correct errors. In the first pass, we were mainly concerned with looking at the phrases over which the events were created and the values of the *Verb* slot in the event. According to the guidelines, events should only be created over either VP chunks (for events centred on verbs) or NP chunks (for events centred on nominalised verbs). In each event, the *Verb* slot should contain *only* the head verb in the phrase (in the case of events created over VP chunks) or the nominalised verb (in the case of events created over NP chunks).

In order to ensure that the final corpus adheres to these rules, the following changes were made:

- Any events created over chunk types other than NP or VP were deleted
- The *Verb* slots of all events were reviewed and edited if necessary so that if the event was created over a VP chunk, only the head verb of the phrase

was included in the *Verb* slot (excluding auxiliary verbs etc), or if the event was created over an NP chunk, only the nominalised verb was included in the *Verb* slot (excluding determiners, adjectives etc).

We also deleted events created over certain verbs which were in our original list of verbs to be annotated, but which we subsequently decided should not be annotated. Annotators were informed of these decisions, but in some cases they continued to annotate them. The verbs were as follows:

- *Have* and *be.* Events expressed by these verbs contain the sort of information that should be found in the ontology (i.e. the *part-of* relation in the case of *have*, or the *is-a* relation for *be).* It was thus decided that there was no need to annotate these verbs.
- Verbs expressing modal information, e.g. the verbs "indicate", "demonstrate", "reveal" etc. when followed by "that", e.g. "The results indicated that ....". According to our modality annotation scheme, such constructions provide modal information about the event that follows, and the verbs within these constructions should not be annotated as event themselves. As these verbs can also be used with non-modal meanings (which we still wished to be annotated), it was necessary to examine the context of each such event within the appropriate abstract in order to determine whether or not it should be deleted.

A final action carried out during this pass through the corpus was to delete any events in which no slots other than the *Verb* slot were filled.

Semantic argument changes

The second pass though the events focussed on the actual semantic arguments of the events, concentrating mainly on whether the forms of the phrases that constitute arguments are appropriate, either according to the general rules for

marking the extent of semantic arguments, or else according to specific guidelines for the semantic role assigned. A small number of changes also involved changing the semantic roles assigned to the arguments. The main changes made are detailed below.

Lists

One of the main items focussed on during the correction of errors in the corpus has been lists of entities. When items in a list correspond to named entities, the guidelines state that the following rules should apply:

- If all items in the list correspond to the same entity type, only one item should be annotated
- If the list contains multiple entity types, then a discontinuous span should be created consisting of one entity of each different type

The main errors found in the annotation of lists are listed below, together with the corrective steps taken:

1)

Error:

A *complete* list of entities is annotated as a continuous span, with a single named entity tag assigned to the whole list.

Solution : We assume that all items in the annotated list belong to the same named entity category. In WordFreak, the extent of the annotated text span is reduced so that it contains only a single item from the list (normally the first item, according to the annotation guidelines).

2)

Error: Multiple items from a list are annotated as part of a discontinuous span, but two or more of the entities within the span are assigned the same named entity category.

Solution: Semantic arguments containing two or more discontinuous spans are clearly marked in the text file of events, with the string “[AND]” occurring between each section of the span. The text span corresponding to the argument is examined in WordFreak and, if the different parts of the span correspond to different items in a list, their named entity categories are examined. If multiple parts of the span are assigned the same category, then parts of the span are removed from the annotation, so that only one item of each named entity category remains.

3)

Error: Individual items in lists are annotated as separate instances of the same semantic role (e.g. multiple THEME roles are created). Multiple instances of a particular role should only be used when two or more arguments of the verb are playing the same role, but are distinct participants in the event, e.g. *X interacts with Y.* In certain cases, such distinct participants can be formulated as a list, e.g. *X and Y interact*. However, in general, different items in lists should not be annotated as separate semantic arguments with the same role name.

Solution: Whenever an event with multiple instances of a particular semantic role is encountered, the text span corresponding to the event is examined in WordFreak. If the items in the separate role instances are individual items from a list, and it is clear that these individual items cannot be considered distinct arguments of the verb, then the spans representing the separate arguments are collapsed into a single (possibly discontinuous) span, so that only one instance of the appropriate role remains within the semantic frame. Certain items contained within the separate role instances may be excluded from this consolidated, single annotation,

according to whether they are assigned the same named entity category as other items in the list.

Other semantic argument errors

Whilst examining the verb frames, a number of other common errors were encountered. These are described below, together with descriptions of actions taken.

1)

Error: According to the annotation guidelines, most semantic arguments should not begin with prepositions (with the exception of *Location* and *Temporal* roles). In addition, certain common phrases that occur at the start of semantic arguments should also be excluded (e.g. *in the presence/absence of* for the CONDITION role). During the examination of the corpus, it was found that in certain cases, annotated semantic arguments began with prepositions or such phrases.

Solution: The extent of the span representing the semantic argument is changed so that the preposition or phrase at the beginning is excluded.

2)

Error: Text spans that are assigned the *Temporal* and *Location* role should include a preposition, if one is present in the text. However, it was found that in certain cases, the preposition was absent from the annotated span even when it is present in the text.

Solution: For each verb frame containing a *Location* or *Temporal* semantic argument, without a preposition at the start of the span, the text span from which the semantic frame was derived is examined in WordFreak. If a preposition precedes the argument in the text, then the text span of the argument is extended

to include the preposition.

3)

Error: The RATE semantic role corresponds to phrases that describe changes in rates or levels that occur as part of the event. As stated in the annotation guidelines, it should *not* be used to annotate phrases that express quantities of other arguments involved in the event. Consider the following sentence: e.g. *Mar mutants of an ompF-lacZ operon fusion strain expressed **50 to 75%** of the beta-galactosidase activity.* Here, *50- 75%* applies to *the beta-galactosidase activity,* rather than expressing a level or rate for the *expressed* event. Marking such phrases as semantic arguments was however a common error found in the annotated corpus.

Solution: The sentence corresponding to each event that includes a *Rate* argument is examined in WordFreak. If the span annotated to represent the argument is clearly a quantification or level of another argument within the frame, then the *Rate* argument is deleted from the frame.

4)

Error: Semantic arguments starting with a temporal preposition such as *during, before* or *after* are marked with the CONDITION role, but should be TEMPORAL

Solution: The assigned role is changed to TEMPORAL

5)

Error: MANNER and INSTRUMENT roles are sometimes confused. According to the guidelines, MANNER phrases correspond to those which describe the *method or way* in which the event occurs or is carried out, whilst the INSTRUMENT semantic role should be assigned to *entities* that are used by the AGENT in order

to carry out the event.

Solution: If the annotated span is clearly an entity, but is annotated with the MANNER role, then the role was changed to INSTRUMENT. Likewise, if a phrase that was clearly a method or action was marked as INSTRUMENT, then it was changed to MANNER.

Correcting syntactic problems

The corpus was found to contain some problems relating to syntax (i.e. chunking). Whilst these problems would not affect the acquisition of semantic event frames from the corpus, it was decided to try to correct the most common types of syntactic problems, in order to ensure more accurate results from the planned syntax-semantic mapping, as well as helping to achieve a more cleanly and accurately annotated corpus, thus enhancing its potential for reusability.

Examination of the corpus revealed two main types of syntactic (i.e. chunking) problems which were introduced as a result of the annotation process. These were as follows:

1) Deeply embedded chunks with the same text span and chunk category. These were very common in the annotated corpus, and were introduced unintentionally by annotators, seemingly as an unexpected consequence of using the “Text” view to carry out annotation within the WordFreak tool. It appeared that every time a text span was selected in this view and used to fill a particular semantic argument slot, a new embedded chunk was automatically created by WordFreak, even when the selected span already corresponded to a syntactic chunk. As the “Text” view does not display embedded chunks, annotators were largely unaware of this. Abstracts

annotated using the “TreeTable” view, where chunks can be selected directly, were not subject to these kinds of problems. A script was run to automatically remove all such embedded chunks. In some cases, other parts of the annotation files had to be updated accordingly, including the ids of event slot fillers, and ids used to form chains (i.e. discontinuous spans).

2) For semantic arguments covering multiple chunks, annotators were asked to assign a type (from a drop-down menu) for a new chunk which would span all chunks in the argument. It was found that frequently, the wrong chunk type was selected by annotators . There seemed to be 2 possible reasons for this:

a) They didn't care or understand properly which chunk type to use, due to their limited linguistic knowledge. Frequently, the chunk type assigned was simply the default one that appeared in the drop-down menu.

b) The semantic argument didn't correspond to one of the "allowed" chunk types for the chosen semantic role. It seems that in some cases, the correct chunk type for the semantic argument was not in the list of foreseen “correct” chunk types for that type of role. In this case, annotators were forced to create a new chunk with an incorrect type. This perhaps suggests that constraining the chunk types of particular arguments in this way was not a good idea. However, by correcting these chunk types, we should get a better idea of the range of chunk types that actually occur for each semantic role type.

A large number of these chunking errors have been corrected, using a combination of automatic and manual processing. The main changes were

as follows:

i) NP changed to PP. There were over 700 cases where multiple-chunk span beginning with a preposition had been assigned the NP tag instead of PP. All of these were corrected automatically. In many cases, it was found that an NE tag had been assigned to the whole span (including the preposition). Normally NE tags should only be assigned to NP chunks. An attempt was thus made to ensure that the NE tag was assigned to the correct NP chunk, within the enclosing PP chunk. Errors were corrected as follows:

a) If the enclosing PP chunk only had one NP daughter, then the NE tag was automatically re-assigned to that NP daughter.
b) All enclosing PP chunks containing 2 or more NP chunks were manually reviewed. In most cases, a new NP chunk was created to enclose all the NP daughters, and the original NE tag to this newly created NP.

ii) NP changed to ADVP. Similar action was taken with these errors as with NPs changed to PPs, as described above.

iii) NP changed to VP-BIO. There were many cases where a VP-BIO was either the single daughter of an NP chunk or an NP was the single daughter of a VP-BIO chunk. Almost all of these were cases corresponded to nominalised verbs over which events had been created. The correct chunking is such cases is for a single (non-embedded) VP-BIO chunk to be assigned, and so all such cases were updated accordingly.

2.2 *Final Annotation results*

When Deliverable D4.1 was released at M20, the linguistic annotation was still ongoing. At that time, 436 single-annotated abstracts had been collected, together with 80 pairs of duplicate-annotated abstracts. Annotation continued until the end of M22, although not all annotators were able to continue beyond M20, or to commit as much time to the task. However, within this extra 2 month period, an extra 161 single-annotated abstracts were produced. Our corpus now contains 697 single-annotated abstracts, in which a total of 3612 separate bio-events have been annotated. The number of duplicate-annotated abstracts remains the same as before, i.e. 80, containing a total of 1158 distinct events.

2.2.1 Corpus Statistics

In D4.1, it was described how abstracts to be annotated were primed by automatically marking all instances from a list of 700 biologically relevant verbs. Only those verbs describing events relevant to gene regulation would be annotated. Our annotated corpus suggests that only a relatively small proportion of these verbs are used to describe such events, at least within abstracts; in total, events were annotated over a total of 277 distinct verbs. Of these verbs, 73 have 10 or more instances annotated in the corpus. Annotators were also instructed to annotate events centred on nominalised verbs, when the nominalised verb occurred as a semantic argument of another annotated event. This has allowed us to identify a total of 135 nominalised verbs that are relevant to the domain, of which 22 have 10 or more instances annotated in our corpus. Table 1 shows the 10 most commonly annotated verbs and nominalised verbs in the corpus, together with the number of times they were annotated, and their type (V=verb, NV= nominalised verb).

Word	Count	Type
Expression	409	NV
Encode	351	V
Transcription	125	NV
Bind	110	V
Require	100	V
Express	93	V
Regulate	91	V
synthesis	90	NV
contain	80	V
induce	78	V

Table 1 Most commonly annotated verbs and nominalised verbs

The fact that 3 out of these top 10 event focus words are nominalised verbs, including the single most commonly annotated word, i.e. *expression,* provides evidence for the assertion that such words play a dominant role in the description of biomedical events (Cohen & Hunter, 2006), and thus emphasises the importance of annotating semantic frame information for them, in addition to verbs.

2.2.1.1 *Semantic Roles*

The counts of semantic roles assigned to arguments of verbs and nominalised verbs in the single-annotator corpus are shown in Table 2. An interesting point to note is that the UNDERSPECIFIED role was assigned only once during the whole annotation project. It will be recalled that this role was made available to assign to semantic arguments whose role did not seem to be well described by one of the other 12 role labels. This suggests that the originally-defined role set has a sufficient scope to describe the vast majority of semantic arguments of gene regulation events, or at least those occurring within abstracts. Although there is a possibility that annotators may have "pigeonholed" certain arguments into potentially unsuitable categories, our review of a large number of annotated abstracts suggests that this is not a common occurrence.

Role Name	Count
THEME	3353
AGENT	1698
LOCATION	526
CONDITION	239
DESCRIPTIVE-THEME	235
MANNER	223
SOURCE	154
DESTINATION	144
DESCRIPTIVE-AGENT	84
RATE	71
INSTRUMENT	60
PURPOSE	57
TEMPORAL	47
UNDERSPECIFIED	1

Table 2: Semantic role counts

The most commonly occurring roles, by a significant margin are THEME and AGENT, which is unsurprising given that these represent “core” event information that must be present (or at least implied) for most events to make sense. It may at first seem surprising that the THEME role is assigned over twice as many times as the agent role. However, this can be best explained by the high occurrence of events described by nominalised verbs, and verbs in the passive form. In these cases, THEMEs are almost always present, but AGENTs rarely so.

It is also interesting to note that 3 out of the next 4 most commonly assigned roles, namely LOCATION, CONDITION and MANNER, correspond to those which Tsai et al. (2007) highlighted as vital for the description of biological events. Our results thus confirm their importance, and reinforce the need for both domain-dependent as well as domain-independent roles within our scheme. The least commonly used roles are INSTRUMENT, PURPOSE and TEMPORAL. However, as minimum number of assignments within the corpus is 47, our results suggest that none of

our defined semantic roles are redundant.

2.2.1.2 *Named Entities*

The single-annotator corpus contains 5401 named entity annotation. All 61 of the defined categories were assigned at least once, with 50 of them being used 10 or more times. The most frequently assigned categories, together with their counts, are shown in Table 3. The most dominant types of entity are thus DNA-based entities, with proteins also being highly pervasive in the description of gene regulation events.

Two of the top ten types correspond to *processes,* rather than entities. Thus, it is highly common for events themselves to form arguments to verbs, a fact which is backed up by the high occurrence of nominalised verbs. The only set of entities that does not figure in the top 10 is the *Experimental* set. This is perhaps to be expected, given that they are most likely to correspond to less commonly occurring role types, such as CONDITION or INSTRUMENT.

Category	**Entity set**	**Count**
GENE	DNA	988
PROTEIN	Protein	602
GENE_ACTIVATION_PATHWAY	Processes	350
ENZYME	Protein	326
PROMOTER	DNA	275
DNA_FRAGMENT	DNA	211
PROKARYOTE_STRAIN	Organisms	191
BIOLOGICAL_PROCESS	Processes	178
OPERON	DNA	155
DNA_STRUCTURE	DNA	130

Table 3 Named Entity Counts

2.2.1.3 *Inter-annotator agreement*

Statistics regarding inter-annotator agreement were presented in Deliverable D4.1. The number of duplicate-annotated abstracts has not increased since then, and so the basic statistics remain the same. However, some more detailed results have since been produced regarding inter-annotator agreement, which are presented in this section. For completeness, the table of inter-annotator agreement results is presented in D4.1 is repeated below in Table 4.

The figures shown in the Table 4 are direct agreement rates. Whilst the Kappa statistic is very familiar in calculating inter-annotator agreement, we follow Wilbur et al. (2006) and Pyysalo (2007) in choosing not to use it, because it does not seem appropriate or possible to calculate it for all of the statistics. For instance:

1. For some tasks, like annotation of events and arguments spans, deciding how to calculate random agreement is not clear
2. The Kappa statistic assumes that annotation categories are discrete and mutually exclusive. This is not the case for the NE categories, which are hierarchically structured.

STATISTIC	**VALUE**
Document pairs	80
EVENTS	
Agreed events	570
Distinct events	1158
Event agreement rate	**0.49**
ARGUMENTS	
Agreed arguments (exact span match)	750
Agreed arguments (partial span match)	912
Distinct arguments	1247

STATISTIC	VALUE
Argument agreement rate (exact span match)	**0.60**
Argument agreement rate (partial span match)	**0.73**
SEMANTIC ROLES	
Agreed semantic roles	712
Semantic role agreement rate	**0.78**
NAMED ENTITIES	
Agreed NEs (exact span match)	502
Agreed NEs (partial span match)	595
Distinct NEs	875
NE agreement rate (exact span match)	**0.57**
NE agreement rate (partial span match)	**0.68**
NAMED ENTITY CATEGORIES	
Agreed NE categories (exact)	374
Agreed NE categories (including parent)	389
Agreed NE categories (including ancestors)	432
NE category agreement rate (exact)	**0.62**
NE category agreement rate (including parent)	**0.65**
NE category agreement rate (including ancestors)	**0.73**

Table 4 Inter-annotator agreement

As described in Table 4, the rate of agreement between identified events is somewhat low, at 49%. However, much discussion was required amongst annotators in order to reach a consensus on the exact nature of the event types to be annotated. Thus, particularly towards the start of the annotation phase, annotators tended to either under- or over-annotate the events, which contributed towards the relatively low agreement figure.

Other parts of the annotation task show higher, and roughly comparable, levels of agreement. The results, do, however, highlight problems on deciding on the exact spans to annotate to represent semantic arguments. Whilst annotators agree in 75% of cases on the number of semantic arguments, and their locations within the sentence, only 60% agreement is reached for exact argument span matches. Despite considerable efforts within our guidelines to enforce consistency of

annotated spans, our results suggest that there may still be some need to refine the guidelines.

Similar levels of agreement were achieved in the identification of NEs. A potential problem here concerned the way in which NEs are annotated within WordFreak. In their normal method of working, annotators had to switch between different views of the text to annotate semantic roles and NEs. Thus, it is possible that annotators sometimes forgot to assign NE categories. For those NEs whose identification was agreed upon, the 62% agreement rate for exact category matches could be increased markedly if cases where the category assigned by one annotator was the ancestor of the category assigned by the other annotator (according to the hierarchical structure of the NEs) were also counted as matches. As there is such a large number of NE categories (i.e. 61), deciding the most appropriate category is often quite a complex task, as verified by annotators in the meetings. Therefore, high rates of agreement on the exact category to assign may be difficult to achieve. However, the hierarchical structure means that it would be possible to use a smaller set of categories by mapping the specific categories to more general ones

Semantic role agreement

The highest rate of agreement is for the assignment of semantic roles, at 78%. However, as shown in Table 4, AGENT and THEME roles make up the vast majority of the semantic roles assigned. We also consider these as the most straightforward of the role labels to assign. Thus, the 78% statistic does not necessarily give a clear indication about how much agreement is reached on the assignment of roles to the less common argument types. We thus calculated agreement rates for the individual semantic roles. These are shown in Table 5, together with a count of the total number of assignments of each role.

Role Name	# of	Agreement

	assignments	
RATE	16	1.00
SOURCE	15	0.93
LOCATION	111	0.90
THEME	975	0.87
AGENT	434	0.85
CONDITION	40	0.80
TEMPORAL	8	0.75
MANNER	59	0.71
PURPOSE	16	0.63
INSTRUMENT	7	0.57
DESTINATION	36	0.44
DESCRIPTIVE-THEME	104	0.46
DESCRIPTIVE-AGENT	35	0.23

Table 5 Agreement rates amongst semantic roles

Whilst the agreement rates for the less commonly occurring roles may not be fully reliable, the table shows that the agreement rates for many of the most frequently occurring roles (i.e. AGENT, THEME, LOCATION, MANNER and CONDITION) lie between 70% and 90%, and thus seem acceptably high.

The 3 roles with the lowest agreement rates all have a reasonable number of occurrences (particularly DESCRIPTIVE-THEME), suggesting that these statistics are fairly accurate. In order to try to understand these rates, we first calculated the types of role disagreements that occur in the corpus. The most common of these are shown in Table 6.

Role 1	Role 2	# of occurences
Theme	Agent	52
Theme	Descriptive-Theme	38
Theme	Descriptive-Agent	20
Theme	Destination	11

Descriptive-Theme	Descriptive-Agent	6
Theme	Manner	6
Purpose	Agent	6
Descriptive-Theme	Agent	5
Location	Destination	5

Table 6 Most common role disagreements

The table shows that the 4 most common disagreements between role assignments all involve the THEME role. Closer examination of the annotated events corresponding to the first 3 types of disagreement reveals that they mainly concern 3 verbs, namely *encode*, *code* and *bind*.

A typical sentence in which disagreement occurs is the following:

***malS**, the gene **encoding** **the periplasmic alpha-amylase**, is under the regulatory control of the MalT protein.*

For such sentences, there are three common patterns of role assignment for the semantic arguments of *encode*:

malS	***The periplasmic alpha-amylase***
AGENT	THEME
THEME	DESCRIPTIVE-THEME
AGENT	DESCRIPTIVE-AGENT

The choice of pattern corresponds to the annotator's interpretation of the event semantics. The AGENT/THEME pattern is most appropriate when the verb describes an action of some kind, whilst the THEME/DESCRIPTIVE-THEME pattern is more suitable when the verb describes a state (i.e. when there no action involved and hence no AGENT). Indeed, the difficulty in annotating *encode, bind* and *code* was discussed during the regular meetings, where it was suggested that

their interpretation can vary according to context, and hence both of these patterns may be appropriate for different occurences of the verbs. However, the fact that there is a fair amount of disagreement of patterns used for particular instances of these verbs suggests the correct interpretation is not always easy to determine, even for domain experts.

For the verb *encode,* a third pattern is observable, namely assigning AGENT to the logical subject of the verb, and DESCRIPTIVE-AGENT to the object. This interpretation suggests that action is involved in the event, but that the subject provides descriptive information about the agent, rather than corresponding to something directly affected by the event. Closer examination showed that this pattern was only used by one annotator. However, it emphasizes the difficulty in correctly categorizing the semantic arguments of this verb in particular.

For the confusions involving the DESTINATION role, the main verb involved is *bind*, as in the following sentence:

In contrast, the ***OmpR2 protein bound*** *preferentially to* ***the ompF promoter***.

The problem again seems to be one of interpretation of the *binding* event. One interpretation is that an AGENT (the subject of the verb) actively binds to a DESTINATION. Another interpretation is that there is no explicit AGENT, and that the entities corresponding to the semantic arguments just happen to bind together, in which case they are both annotated as THEMEs. Table 7 shows that a second type of confusion is between DESTINATION and LOCATION. This is an understandable confusion, as both roles correspond to locative information.

2.2.1.4 *Summary of results*

The results presented in the above sections suggest that the annotation scheme is well suited to describing events within the gene regulation domain. On the one

hand, all roles within in the scheme were assigned a sufficient number of times provide evidence of their usefulness. On the other hand, it seems that our proposed role set is sufficiently general and wide-ranging to characterize the vast majority of semantic arguments of gene regulation events. Furthermore, our inter-annotator agreement rates suggest that semantic arguments can be identified and categorized fairly consistently by different annotators. Where disagreements did occur, these were found to be concentrated on a relatively small number of verbs, with fairly regular alternations of role assignment patterns.

Examination of our annotated corpus has, however, identified a number of problematic areas. These include the choice of which verbs to annotate as gene regulation events, which exact text spans to annotate to represent semantic arguments and the choice of the most appropriate named entity categories. Whilst guidance relating to all of these is provided in the annotation guidelines, the lower inter-annotator agreement rates for these annotation subtasks suggest that the guidelines may benefit from some revision prior to carrying out any subsequent annotation based on this scheme. Furthermore, the higher agreement rates achieved when considering higher-level named entity categories suggests that using a more coarse-grained set of categories should perhaps be considered if further annotation is carried out.

2.3 *Modality annotation*

In this section, the results of the feasibility study carried out for annotating modality information on previously annotated bio-events is reported.

2.3.1 Testing the classification scheme

The feasibility study consisted of the annotation of modality on a small set of 202

abstracts, that had previously annotated with BELA events. The annotation was carried out using *WordFreak.* Due to the linguistically-driven purposes as well as the small size of the corpus exploited, annotation was carried out by a single annotator at CNR-ILC with linguistic expertise. However, extensive support was provided by two BOOTStrep team members, one with a background in linguistics, and the other one in biology, to discuss open issues raised during the annotation process in order to improve the semantic stability and reliability of the annotations produced.

Each sentence containing a previously-annotated gene regulation event was studied, and modality annotation was performed only on those sentences in which the description of the event contained explicit expression of modal information: modal information was only annotated if it was within the scope of the gene regulation event described. Let us consider, for example, the *derepress* bio-event, described in the sentence "*We suggest that overproduction of SlyA in hns(+) E. coli derepresses clyA transcription by counteracting H-NS"*, which was annotated as follows:

VERB: *derepresses*
AGENT: *overproduction*
THEME: *clyA transcription*
MANNER: *counteracting*

The modality annotation process started from the event anchor, i.e. the verb *derepress*. Words or phrases expressing modal information and linguistically bound to the event anchor were searched for within the sentence's span. If such items were found, values from the proposed sets were selected for one or more of the three dimensions of the annotation scheme, i.e. *Point of View, Knowledge Type* and *Certainty Level*. For the *Knowledge Type* and *Certainty Level* attributes,

a value was only selected if there was *explicit* lexical evidence in the sentence. In the case at hand, *suggest* was annotated as the lexical modality marker conveying information about *Knowledge Type*, whose associated value is *deductive*. The word *We* was interpreted as lexical evidence that the reported *Point Of View* was that of the writer.

Each piece of lexical evidence (i.e. lexical modality marker) could only be used to assign a value to *one* of the annotation dimensions. Thus, it was not possible to use a single word or phrase to assign values to both the *Knowledge Type* and *Certainty Level* dimensions.

If one or both the *Knowledge Type* or *Certainty Level* attributes were assigned, the *Point of View* attribute was also instantiated. If no explicit lexical evidence was available for the assignment of this attribute, a "default" value of *writer* was assigned, i.e. it was assumed that the *Point of View* was expressed implicitly.

The annotator used the preliminary categorisation of modal lexical items as a starting point for the annotation of the *Knowledge Type* and *Certainty Level* attributes, although she was not bound by this categorisation, nor was her annotation limited to only those items on the list: part of the purpose of the annotation was to discover the semantic stability of the lexical items within our proposed categories, as well as to discover other modality markers missing from the preliminary list.

2.3.2 Results

The 202 MEDLINE abstracts annotated for modal information contained a total of 1469 gene regulation events. 249 of these events (i.e. 16.95%) were annotated with modality information. Table 7 shows general statistics about the *dimensions* of

the modal markers that were present in the description these events, whilst Table 8 shows the distribution of the annotations amongst the various values within each dimension of the scheme.

Modal marker(s) present	**Count**	**% of total events**
Knowledge Type only	192	77.11 %
Certainty Level only	40	16.07%
Knowledge Type + Certainty Level	17	6.83%

Table 7. Distribution of modality markers within annotated events

Dimension	**Value**	**Count**	**% of annotations within dimension**
Knowledge Type	DEMONSTRATIV	110	52.63%
	DEDUCTIVE	56	26.79%
	SENSORY	25	11.96%
	SPECULATIVE	18	8.61%
Certainty Level	ABSOLUTE	4	7.01%
	HIGH	15	26.31%
	MODERATE	34	59.64%
	LOW	2	3.50%
Point Of View	WRITER	213	92.20%
	OTHER	18	7.79%

Table 8. Distribution of modality annotations within the different dimensions

The number of modality annotations may at first seem rather low, with an average of 1.31 annotations per abstract. However, a number of points should be noted. Firstly, lexical markers of modality are generally quite sparse within texts. Secondly, as pointed out above, modality annotations have only been carried out on top of previously annotated bio-events, and there was an average of 6.05

bio-event annotations per abstract. Rather than aiming to annotate *all* modal information expressed within the abstracts, our case study is firstly aimed at verifying whether the modality classification scheme is suitable for a corpus of biomedical texts, and secondly, it is focused on the discovery of the main domain-relevant problems and features involved, as well as clues which can drive future work.

There follows a number of annotation examples. In each case, the modality marker(s) and the Point Of View marker (if present) have been underlined, with the corresponding category placed in brackets. The verb which forms the focus of the associated bio-event is emboldened.

a) *Therefore, we* [WRITER] *suggest* [DEDUCTIVE] *that overproduction of SlyA in hns(+) E. coli* ***derepresses*** *clyA transcription by counteracting H-NS.*
b)*We* [WRITER] *have shown* [DEMONSTRATIVE] *that the open reading frame ybbI in the genomic sequence of Escherichia coli K-12* ***encodes*** *the regulator of expression of the copper-exporting ATPase, CopA*
c)*We* [WRITER] *speculate* [SPECULATIVE] *that the product of this gene is* ***involved*** *in the attachment of phosphate or phosphorylethanolamine to the core and that it is the lack of one of these substituents which results in the deep rough phenotype.*

A single modality marker may also express the same information relative to more than one bio-event in the case of a coordinated structure, e.g. :

Band shift experiments showed [DEMONSTRATIVE] *that AllR* ***binds*** *to DNA containing the allS-allA intergenic region and the gcl(P) promoter and its binding is* ***abolished*** *by glyoxylate.*

2.3.2.1 *Knowledge Type* information

The *Knowledge Type* dimension is the most frequently annotated (77.11% of annotations). The most common value for this dimension is *demonstrative* (52.63% of Knowledge Type annotations), whilst the least widespread type of knowledge is *speculative* (8.61%).

These statistics are perhaps unsurprising, given that the current pilot study has been carried out on abstracts. *Demonstrative* events are explicitly marked as describing experimental results, particularly those which prove hypotheses or predictions. These are exactly the sorts of events that we can expect to occur most frequently in abstracts; within the short amount of space available, authors normally aim to emphasize the definite results that their experiments have produced.

The annotation experiment has highlighted a potential need to add an additional value for the *Knowledge Type* dimension. Consider the following examples:

*a) The model states that the lex (or exrA in E. coli B) gene **codes** for a repressor*.
b) *Mutations in yjfQ allowed us to identify this gene as the regulator of the operon yjfS-X (ula operon), reported to be **involved** in L-ascorbate metabolism.*

Events that are introduced by verbs such as *state* or *report* do not fit well into one of our other four *Knowledge Type* categories. They are used to introduce facts, either cited from previous work or earlier in the paper, but without taking a particular stance to them, i.e. there is no speculation or deduction involved, and there is no reference to active proof or demonstration that an assertion or hypothesis is true.

Statements such as the above fit into Hyland's (1996a) description of the quotative

category, i.e. specifying and acknowledging previous findings. Thus, the quotative label can apply to a wider range of statements than just those that contain citations. Therefore, we propose to introduce the quotative category into our classification as a further *Knowledge Type* category to cover statements that specify or acknowledge previous findings through explicit lexical items.

Our annotation also revealed that, whilst the majority of *Knowledge Type* items are fairly stable semantically within their assigned categories, a small number of items do not fit neatly within a single category. The verb *seem* was originally placed within the *sensory* category, following Hyland. However, there is often a speculative aspect to its meaning, as confirmed by Dixon (2005): *seem* is used "when there is not quite enough evidence" (p. 205). The degree of speculation conveyed may vary according to the context: this is an area for further research.

2.3.2.2 *Certainty* information

Certainty level markers are considerably less common than *Knowledge Type* markers, representing 16.07% of the modality annotations. The most widespread value among these annotations is *moderate* (59.64%).

The high percentage of *moderate* markers can again be explained by the text type, i.e. abstracts. The results concerning *Knowledge Type* illustrated that *demonstrative* statements are most common: authors are keen to emphasize the experimental results that they have produced. If there is doubt about these results, this can be indicated thought an explicit certainty level marker. A *moderate* (and hence neutral) certainty level marker may be the "safest" choice here.

Certainty Level markers occur most commonly without an accompanying *Knoweldge Type* marker, as in:

*EvgA is likely [HIGH] to directly **upregulate** operons in the first class, and indirectly upregulate operons in the second class via YdeO.*

As mentioned previously, *Knowledge Type* markers implicitly encode certainty level information. Thus, when a statement is explicitly marked as a speculation or deduction, the use of an explicit marker of certainty may be unnecessary, except for emphasis, or to alter the "default" certainty level associated with the *Knowledge Type* item.

Nevertheless, our annotation has served to identify a small number of cases (6.83%) that contain explicit markers of both *Knowledge Type* and *Certainty Level* information. Such cases provide evidence that our proposed separate dimensions of annotation are indeed well motivated. Some examples are shown below:

a) *No reverse transcriptase PCR product could be detected for hyfJ-hyfR, suggesting* [DEDUCTIVE] *that hyfR-focB may* [MODERATE] *be independently **transcribed** from the rest of the hyf operon.*
b) *We* [WRITER] *suggest* [SPECULATIVE] *that these two proteins may* [MODERATE] ***form** a complex in the membrane which acts at late steps in the export process.*

A large number of certainty level markers are fairly stable in terms of semantics, particularly adjectives and adverbs such as *probable, possibly* or *likely.* Another category of words that play a central role in expressing certainty in our corpus is the modal auxiliaries (e.g. *can, may* or *could*), which represent 40.35% of the total number of *Certainty Level* markers. However, their interpretation is more problematic than adjectives and adverbs like those listed above. In general, *can, may* and *could* can have the following senses:

a) *Moderate* level of certainty
b) Theoretical possibility (indicating that an event has the potential to occur)
c) Ability
d) Permission

Whilst the *permission* sense is rarely relevant within biomedical texts, examples of the other three senses can be readily identified within our corpus. Some examples involving *may* are shown below:

1) **Certainty level marker**
The DNA-binding properties of mutations at positions 849 and 668 may [MODERATE] *indicate* [DEDUCTIVE] *that the catalytic role of these side chains is **associated** with their interaction with the DNA substrate.*
2) **Theoretical possibility marker**
*The expression of nifC may be **coregulated** with nitrogen fixation because of the presence of nif-distinctive promoter and upstream sequences preceding nifC-nifV omega-nifV alpha.*
3) **Ability marker**
*Results obtained indicate that the nrdB gene has a promoter from which it may be **transcribed** independently of the nrdA gene.*

Thus, the presence of these modal auxiliaries does not guarantee that certainty level is being conveyed. Determining the correct sense can be a difficult task, which requires in-depth knowledge of the domain, and often requires examining a wider context than just the sentence itself.

Whilst this could prove problematic in the automatic recognition of modality, Collins (2006) suggests that for each verb, one sense is usually more likely than the

others. In his study of *can* and *may* in various spoken and written sources, he found that *may* was used as a certainty level marker in 83.5% of cases, whilst only 1.1% of occurrences of *can* concerned certainty level. A default interpretation of each modal could thus be used. Further study of the context of these items may reveal clues that could determine when a non-default value should be assigned.

Our studies have shown that the meaning of *can* mainly corresponds to the "ability" sense, although "theoretical possibility" is also possible, as shown in the following examples:

a) *The enhanced expression of tac-dnaQ reduces 10-fold the frequency of UV-induced Su+ (GAG) mutations in the CCC phage and nearly completely prevents generation by UV of Su+ (GAG) mutations in the GGG phage, in which UV-induced pyrimidine photo-products can be* ***formed*** *only in the vicinity of the target triplet.*
b) *These results indicate that OmpR stabilizes the formation of an RNA polymerase-promoter complex, possibly a closed promoter complex, and that a transcription activator can* ***serve*** *not only as a positive but also as a negative regulator for gene expression.*

Whilst the "ability" sense is not central to the interpretation of modality, the recognition of "theoretical possibility" may be more important: stating that an event has the *potential* to happen is different from stating that it *does* (always) happen. Thus, further investigation of lexical markers of theoretical possibility will help to build upon our current categorisation model.

2.3.2.3 *Point Of View* information

Although we suggested that there are a number of textual clues that can be used to

determine the *Point of View* of a statement, our annotation experiment revealed that such explicit evidence is quite sparse, at least in abstracts. Occasionally, the sentence contains words or phrases such as *we*, *our results*, *in this study*, etc. allowing the *Point Of View* to be determined as the author(s) of the abstract. In other cases, looking at the wider surrounding context, i.e. in neighbouring sentences or even within the whole abstract, is necessary. Although our annotation assumes the lack of an explicit *Point of View* marker to indicate the *writer* point of view, further analysis of these cases must be carried out.

During annotation, however, we identified some potential additional clues that can help to determine the value of this dimension.

Consider the phrase *these results*. On its own, this provides no explicit information about the point of view of the accompanying statement. However, when occurring as the subject of *suggest* (especially in the present tense), it is normally the case that the deduction has been carried out by the author(s), as illustrated in the following example:

These results [WRITER] *suggest* [DEDUCTIVE] *that both locally and regionally targeted mutagenesis is* ***affected*** *by overproduction of the epsilon subunit.*

The *writer* value can also be assumed in such contexts when other verbs in the *deductive* and *sensory* categories are used, e.g. *indicate, imply, appear*, etc, particularly when in the present tense with an inanimate subject. An exception is when there is explicit reference to another author or work. If there is an impersonal subject, e.g. *It is suggested*, then greater contextual evidence would be required, as the point of view is ambiguous.
A further example concerns *Certainty Level* markers within the *absolute* category, which generally denote well-established facts within the community. When such a

certainty level marker is present, we can assume that the statement does not correspond only to the author's personal point of view. An example is shown below:

Near the amino terminus is the sequence 35GLSGSGKS, which exemplifies a motif known [ABSOLUTE] *to* ***interact*** *with the beta-phosphoryl group of purine nucleotides.*

2.3.3 Summary of modality results

In many cases, textual clues can be used fairly reliably to determine the correct classification of statements according to the dimensions of *Knowledge Type, Certainty Level* and *Point of View*. The results from a preliminary annotation experiment based on this scheme confirm this hypothesis.

Contextual information surrounding modal lexical items can also be important in determining the correct modal value of statements. Shallow parsing (i.e. chunking), on the top of which event annotation and modality annotation are carried out, can help to identify such information. This is in agreement with Medlock & Briscoe (2006), who suggest that linguistically-motivated knowledge may help to boost the performance of an automatic hedge classification system.

Our preliminary results suggest that many modal items in our list are fairy stable semantically when modifying bio-events. However, the correct interpretation of modal auxiliaries within the domain is more problematic, and is thus an area for further research. Our experiment also served to highlight certain weaknesses in the original model, e.g. the lack of a *Knowledge Type* category corresponding to reported facts. A further potential weakness in our results is that, whilst examples

supporting all of our proposed categories were found, there is a strong bias towards certain categories. However, this may be because our preliminary study was based only on abstracts.

3 Extraction of bio-event frames from the BELA corpus

The bio-event linguistically annotated corpus is currently being used to train and test information extraction methods that acquire bio-event frames to be used for populating the Bio-Lexicon. In this section, the extraction methodology and achieved results will be illustrated in detail.

3.1 *Event Patterns*

Event patterns are fragments of event annotations in which semantic arguments are generalized to their semantic role and named entity categories, if present.

An event pattern is extracted for each unique event id within an abstract. An event annotation span begins with the earliest SLOT span, and ends with the latest SPAN assigned to the event. An example event span is as follows:

<SLOT eventid="9" Role="Agent"> <NE cat="OPERON"> transfer operon</NE></SLOT> <EVENT id="9"><SLOT eventid="9" Role="Verb"> expression </SLOT></EVENT></SLOT> of <SLOT eventid="9" Role="Theme"> <NE cat="DNA_FRAGMENT"> F-like plasmids </NE></SLOT>

Event spans are generalized into event patterns as follows:

“Verb” role slots are converted into a string consisting of the role type, part-of-speech and surface form, i.e., [Verb:*POS*:*verb*].

Word sequences annotated with other semantic role types and/or named entity

tags are generalized to the role and/or named entity super class, i.e., [*role*:*NE_super_class*].
Other XML tags are removed.

The above example event span is thus generalized to the following event pattern:

[Agent:DNA] [Verb:NN:expression] of [Theme:DNA].

3.2 *Event frames*

Event frames are directly extracted from event patterns, and take the following general form:

```
event_frame_name(
     slot_name => slot_value,
     …
     slot_name => slot_value).
```

where event_frame_name is the base form of the event verb or nominalized verb; slot_names are the names of the semantic roles within the event pattern; slot_values are named entity categories, if present within the event pattern.

For example, the event frame corresponding to the event pattern shown in the previous section is as follows:

```
expression( Agent=>DNA,
            Theme=>DNA ).
```

Note that event frames are independent of the original surface syntactic patterns.

3.3 *Corpus Format*

For the purposes of event frame extraction, the annotations in the corpus were converted to an XML-style inline format, consisting of three different types of element:

1) **EVENT-** surrounds text spans (typically verb phrases and nominalised verbs) on which events are centred.

Attributes:

- **id –** a unique id for the event

2) **SLOT** – surrounds spans corresponding to semantic arguments (or slots) of events.

Attributes:

- **argid –** a unique id for the semantic argument. Note that particular semantic arguments can correspond to multiple, discontinous spans of text. In such cases, the same "argid" is assigned to each part of the argument.
- **eventid –** the id of the event to which the argument belongs
- **role –** the semantic role assigned to this argument

3) **NE –** surrounds text spans annotated as named entities.

Attributes:

- **cat –** the category assigned to the NE

In the case that there are several annotations over a particular text span, then elements are embedded inside each other. If more than one annotation begins at a particular offset, then the ordering of the embedding is fixed, so that SLOT elements are embedded inside EVENT elements, and NE elements are embedded inside SLOT elements. An example of the annotation for the sentence

"*TaqI restriction endonuclease has been subcloned downstream from an inducible phoA promoter*" is shown below:

<SLOT argid="4" eventid="5" Role="Theme"> <NE cat="ENZYME">TaqI restriction endonuclease</NE></SLOT> <EVENT id="5">has been <SLOT argid="6" eventid="5" Role="Verb">subcloned </SLOT></EVENT> <SLOT argid="8" eventid="5" Role="Location">downstream from <NE cat="PROMOTER">an inducible phoA promoter</NE></SLOT>.

The EVENT created over the VP chunk *has been subcloned* has been annotated as having 2 semantic arguments (SLOTs), i.e. a THEME, *TaqI restriction endonuclease* and a LOCATION, i.e. *downstream from an inducible phoA promoter.* A 3rd SLOT element (with the role type VERB) corresponds to the head verb in the VP chunk. Named entity tags have also been assigned to the THEME span and part of the LOCATION span.

3.4 *Event Frame Extraction*

Event frame extraction is a fusion of sequential labelling based on conditional random fields (CRF), and event pattern matching. Event frames are extracted in three steps. Firstly, a CRF-based named entity recognizer (NER) assigns biological named entities to word sequences. Secondly, a CRF-based semantic role labeller determines the semantic roles of word sequences with NE labels. Thirdly, word sequences are compared with event patterns derived from the corpus. Only those event frames whose semantic roles, NEs, and verb POS satisfy event pattern conditions will be extracted.

3.5 *NER*

Since it is inherently costly and time consuming to create a large-scale training corpus annotated by biologists, we need to concede to use coarse-grained biological NE categories. That is, the NER component is trained on the five NE super classes, i.e., *Protein, DNA, Experimental, Organisms*, and *Processes*.

The NER models are trained by CRFs (Lafferty et al., 2001) using standard IOB2 labeling method. That is, the label ``B-*NE*'' is given to the first token of the target *NE* sequence, "I-*NE*" to each remaining token in the target sequence, and ``O'' to other tokens.

Features used are as follows:

- word feature
  - orthographic features:
  - ➢ the first letter and the last four letters of the word form, in which capital letters in a word are normalized to "A", lower case letters are normalized to "a", and digits are replaced by "0". For example, the word form "IL-2" is normalised to "AA-0".
  - postfix features: the last two and four letters
- POS feature

We applied first-order CRFs using the above features for the tokens within a window size of ±2 of the current token.

3.6 *Semantic Role Labeling*

First of all, each NE token sequence identified by B and I labels is merged into a single token with the NE category name. Then, the semantic role labelling models are trained by CRFs in a similar way to NER. That is, the label ``B-*Role*'' is given to the first token of the target *Role* sequence, "I-*Role*" to each remaining token in the target sequence, and "O" to other tokens.

Features used here are as follows:

- word feature
- base form feature
- POS feature
- NE feature

The window size was set to ±2 of the current token.

3.7 *Event pattern matching*

Token sequences with NE and semantic role labels are compared with event patterns. The token sequences are converted into annotated sentences in the following manner:

Each token sequence labelled by IOB semantic role labels is merged into a token labelled with the role.

- Verbs and nominalized verbs are converted to [*Verb*:*POS*:*surface_form*].

- Semantic role label, NE super-class, and surface token are converted into the form [*Role*:*NE_super_class*].

- Other tokens with O label are converted to surface tokens.

Then, event patterns are generalized:

Event patterns are modified so that elements corresponding to verbs and

nominalized verbs will match any words with the same POS, e.g., [*Verb*:*POS*:*].

Finally, each event pattern is applied to annotated sentences one by one:

By matching the generalized event patterns with annotated token sequences, i.e. when verbs or nominalized verbs and the surrounding semantic roles and NEs satisfy the event pattern conditions, then successfully unified event patterns are extracted as new event patterns.
The newly obtained event patterns are converted into event frames in the same way as described in Section 4.1.

3.8 *Experimental Results*

The aim of this section is to evaluate semantic frame extraction performance, given a set of annotated training data.

The annotated corpus was randomly separated into 10 document groups and their event patterns and event frames were segmented into 10 groups according to the document separation.

We conducted 10-fold cross validation based on the 10 document groups. Named entity recognizers and semantic role labelers are trained using 9 groups of annotated documents. The results are evaluated on the remaining group of documents and their case frames. We extracted 885 distinct event frames from the corpus.

Table 9 shows the event frame extraction performance for each fold. #TP, #FN, and #FP indicate the number of true positives, false negatives, and false positives,

respectively.

Named entity recognition performance is also evaluated (Table 10). Since the training data size is small, the performance is between approximately 20-60% F-measure. However, this will not cause a problem for the event frame extraction task. This is because, if a particular event frame occurs multiple times in a corpus, it is sufficient to extract only a single occurrence of the event description. So, whilst the NE and semantic role labelling may not be successful for all occurrences of the event frame, there is a good chance that at least one occurrence of the event will be realized in the text in such a way as to allow the labeling to be carried out successfully, thus allowing the extraction of an appropriate event frame.

	Score	#TP	#FN	#FP
Recall	0.186	165	730	
Precision	0.490	165		172

Table 9. 10-fold cross validation results

NE Type	Recall	Precision	F
DNA	0.627	0.660	0.643
Protein	0.525	0.633	0.574
Experimental	0.224	0.512	0.312
Processes	0.125	0.337	0.182
Organisms	0.412	0.599	0.488

Table 10. NE identification performance

4 Representation of acquired bio-event frames in the Bio-Lexicon

In this section, the representation of acquired bio-event frames in the Bio-Lexicon will be discussed, with a specific view to the target representation and the

interchange format to be used for uploading acquired information into the Bio-Lexicon.

Event frames are encoded in the following XML format.

```
<Cluster CLSID="CLSID" SEMTYPE="General">
    <Entry entryid="EntryID" BASEFORM="BaseForm" type="PREFERRED">

      <SemanticPredicate id="SemanticPredicateID">
         <SemanticArgument id="SemanticArgumentID">
            <feat att="role" val="Role"></feat>
            <feat att="restriction" val="NE"></feat>
         </SemanticArgument>
          ...
         <SemanticArgument id="SemanticArgumentID">
          ...
         </SemanticArgument>
      </SemanticPredicate>
        ...
      <SemanticPredicate id="SemanticPredicateID">
        ...
      </SemanticPredicate>
   </Entry>
</Cluster>
```

```
-------------------------------------------------------------------
----
    SemanticPredicateID ::= MANCU_Verb_Number
    SemanticArgumentID ::= SemanticPredicateID_Role

Role ::= agent|theme|manner|instrument|location|source|destination|
temporal|condition|rate|descriptive-agent|descriptive-theme|purpose
```

```
    NE ::= Protein|DNA|Experimantal|Organisms|Processes

-------------------------------------------------------------------
----
```

For example, "interact(Agent=>Protein)" and "interact(Agent=>Protein, Theme=>DNA)" are presented as follows.

```
<Cluster CLSID="MANCU_V1IN224">
<Entry entryid="MANCU_V1IN224_1" baseform="interact" type="PREFERRED">

      <SemanticPredicate id="MANCU_Interact_1">
         <SemanticArgument id="MANCU_Interact_1_agent">
           <feat att="role" val="Agent"></feat>
           <feat att="restriction" val="Protein"></feat>
         </SemanticArgument>
      </SemanticPredicate>

      <SemanticPredicate id="MANCU_Interact_2">
         <SemanticArgument id="MANCU_Interact_2_agent">
           <feat att="role" val="Agent"></feat>
           <feat att="restriction" val="Protein"></feat>
         </SemanticArgument>
         <SemanticArgument id="MANCU_Interact_2_theme">
           <feat att="role" val="Theme"></feat>
           <feat att="restriction" val="DNA"></feat>
         </SemanticArgument>
      </SemanticPredicate>
   </Entry>
</Cluster>
```

5 Syntax-semantics linking

The possibility of linking the subcategorisation frames associated with biologically relevant verbs with extracted Bio-event frames was explored in cooperation with WP3. In spite of the fact that the issue of the syntax-semantics linking was not originally foreseen in the Technical Annex, we believe that acquired subcategorisation frames should be linked to corresponding bio-event frames in the BL.

It is well known that Information Extraction applications require sophisticated lexical resources to support their processing goals. In particular, accurate applications focused on extraction of event information from texts require resources providing an exhaustive account of the semantic and syntactic combinatorial properties of lexical units conveying event information. We have seen that both syntactic subcategorization and semantic event frames have been acquired within the project from a biomedical corpus on the subject of E. Coli. However, the two sets of frames were obtained independently, using different techniques operating on corpora of different size annotated with different information types and resulting in two different and unrelated sets of subcategorization and semantic event frames. For these information types to be exploited more effectively in IE applications, the syntax-semantics linking was performed manually on the two sets of frames acquired for the same verbs; in this section the linking process is described in detail. We believe that a lexicon including subcategorization and semantic frames information as well as the explicit linking between semantic and syntactic slots in corresponding frames has the potential to effectively support event extraction from biomedical texts.

5.1 *The starting point*

The starting point of the syntax-semantics linking was represented by:

1. the list of 856 verbal Bio-event frames extracted from the BELA corpus (see section 3), as exemplified below for the verb *abolish*:
   - eframe('verb'=>'abolish', 'Agent'=>'DNA', 'Theme'=>'O')
   - eframe('verb'=>'abolish', 'Agent'=>'DNA', 'Theme'=>'O', 'Manner'=>'O')
   - eframe('verb'=>'abolish', 'Agent'=>'DNA', 'Theme'=>'Protein', 'Location'=>'O')
   - eframe('verb'=>'abolish', 'Agent'=>'O', 'Theme'=>'O')

   Note that the list of frames includes slots which are assigned a named entity category as well as slots which are not specified for this information type.
2. the list of 1760 subcategorization frames, acquired from the Enju annotated corpus (see Deliverable 3.3, section 4), as exemplified in the table below:

Verb	DEP_1	DEP_2	DEP_3	All dep	subcat freq	p(subcat|v)	Pass
Abolish	ARG1	ARG2		1075	932	0.8669767	0.1437768
Abolish	ARG1	ARG2	MOD@VBG	1075	42	0.0390697	0.1904761
Abolish	ARG1	ARG2	PP-in	1075	101	0.0939534	0.7029702
accumulate	ARG1	ARG2		1180	347	0.2940677	0.0403458
accumulate	ARG1			1180	546	0.4627118	0
accumulate	ARG1	ARG2	PP-in	1180	128	0.1084745	0.140625
accumulate	ARG1	PP-in		1180	159	0.1347457	0

It turned out that for 168 verbs both subcategorization and event frame information was available. For this linking experiment we focused on this subset, in particular on the 628 subcategorization frames and the 486 event frames automatically extracted for these verbs respectively from the Enju annotated corpus and the BELA corpus; note that for event frames, abstraction was made from the NE categorisation of individual slot fillers.

5.2 *Approaching the problem*

In defining our approach to the syntax-semantics linking, different issues were taken into account, in particular:

- the fact widely acknowledged in the linguistic literature that the syntactic realisation of semantic arguments is not accidental, e.g.:
    - "Agents" are typically expressed as subjects in English,
    - "Patients" can be either subjects or objects, as for example in *John* ("Agent") *opened the door* ("Patient") vs *The door*("Patient") *opened*;
- the fact that there are systematic alternations in the syntactic expression of verbal arguments (so-called diathesis alternations), e.g.:
    - at the level of grammar, as observed by comparing the active and the passive voice, for example in *John gave Mary a book* vs *Mary was given a book*;
    - at the level of individual lexical items or classes of them, as in the case of *Peter sprayed water on his flowers* vs *Peter sprayed his flowers with water*.

The linking between extracted subcategorization and bio-event frames was defined by combining different information types, namely:

- we considered that a syntax-semantic mapping process is controlled by strategies which presuppose hierarchies of thematic roles and grammatical functions;
- we resorted to a list of 'prototypic' syntactic realisations of semantic arguments, as fixed in the Annotation Guidelines followed by annotators during the manual annotation of bio-event frames in the selected domain corpora;
- we exploited general language repositories of semantic frames (e.g. VerbNet) containing both syntactic and semantic information as possible benchmarks,

- we also resorted to the manually annotated BELA corpus, when the evidence of the other information sources was not sufficient to perform the syntax-semantics mapping.

In what follows, the different information types which have been used to drive the linking process are discussed in detail.

5.2.1 Analysis of the syntax-semantics linking literature

Firstly, we analysed the syntax-semantics linking literature according to which "Thematic Hierarchies" (henceforth, TH) appear to be by far the most widely used method to explain the mapping from semantic representation to syntax. Fillmore (1968) was the first to formulate a hierarchy of "cases" (semantic relations) to help determine subject selection. After him, most theories make use of a mapping between an ordered list of semantic roles and an ordered list of grammatical relations. Thus, rather than having invariable correspondence relations, these approaches suggest that, given a thematic role hierarchy (agent>theme ...) and a syntactic functions hierarchy (subject>object ...), the mapping usually proceeds from left to right, mapping the semantic role further to the left onto the first available position in the syntactic hierarchy.

Several proposals have been made for what concerns the thematic role hierarchy which widely differ a) with respect to the theoretical stands and b) in what is being hierarchisized. If on the one hand there is general agreement on the fact that the Agent role should be the highest ranking role, on the other hand no consensus is found for what concerns the relative ordering of the remaining roles. To illustrate the wide range of proposals put forward in the literature, consider the following ones collected by Levin and Rappaport (1996):

- TH with no mention of goal and location:

  - Belletti & Rizzi 1988: Agt > Exp > Th
  - Fillmore 1968: Agt > Inst > Obj
- TH with goal and location ranked above theme/patient:
  - Grimshaw 1990: Agt > Exp > G/S/L > Th
  - Jackendoff 1972: Agt G/S/L > Th
  - Van Valin 1990 Agt > Eff > Exp > L > Th > Pat
- TH with goal and location ranked below theme/patient:
  - Speas 1990: Agt > Exp > Th > G/S/L > Man./Time
  - Carrier-Duncan 1985: Agt > Th > G/S/L
  - Jackendoff 1990: Act > Pat/Ben > Th > G/S/L
  - Larson 1988: Agt > Th > G > Obl
  - Baker 1989: Agt > Inst > Th/Pat > G/L
- TH with goal above patient/theme; location ranked below theme/patient:
  - Bresnan & Kanerva 1989: Agt> Ben > Rec/Exp > Inst> Th/Pat>L
  - Givón 1984: Agt > Dat/Ben > Pat > L > Inst

It should be noticed that most of the disagreement lies in where to locate the Theme with respect to other roles, especially the Goal and Location roles. Two reasons explain the difficulty in locating the Theme. First, Theme/Patient arguments can be both subjects and objects. The second reason is that the Theme/Patient competes with the Goal argument to be the first object of verbs taking double objects.

A widely accepted syntactic functions hierarchy is reported below:

- subject > object > indirect object > oblique.

Besides differences in the inventory and relative ordering of thematic roles in TH, a widely shared assumption is that semantics-to-syntax mapping preserves prominence relations between arguments. Note that prominence preservation

does not require that an argument bearing a particular semantic role have a unique syntactic realization, but that each asymmetric relation in the semantic representation is mapped onto a similarly asymmetric relation in the syntax. To be more concrete, either agents or themes may be realized as subjects, as long as their coargument is lower ranked; similarly, either themes or locations may be realized as objects, as long as their coargument is higher ranked.

Such a prominence preserving constraint was resorted to to guide the syntax-semantic linking of slots in the extracted subcategorization and bio-event frames. It was not to be taken for granted that such a constraint could be applied extensively to domain-specific bio-event annotations, also including a set of domain-specific roles.

5.2.2 A list of 'prototypic' syntactic realisations of semantic arguments

Another important source of information was represented by the 'prototypic' syntactic realisations of semantic arguments as defined in the BELA annotation Guidelines (see Deliverable 4.1, Appendix 1), especially for what concerns less prominent roles, typically expressed as prepositional phrases. In the Guidelines provided to the annotators, it was explicitly stated whether a semantic role filler could be introduced by a preposition and, if it was the case, the prototypical preposition types were also specified. The following table summarises the Annotation Guidelines for what concerns the prototypical syntactic realisation of semantic role fillers:

Semantic role	Preposition	Preposition type	Main annotation guidelines
AGENT	-		It can occur as the **subject** of the verb
	+		It can occur in positions other than the subject of the verb, e.g. in a sentence such as "X *results* from Y", where Y is the Agent of the verb *result*
	+	by	In passive sentences, if an agent is present it follows the verb and it is

Semantic role	Preposition	Preposition type	Main annotation guidelines
			preceded by the preposition “by”
			In passive sentences it can be omitted
	-		It is possible for an event to have more than one Agent, if more than one of the variables in the event can be considered to be responsible for causing the event
THEME	-		It mostly occurs as the **object** of the verb
	-		It can occur in positions other than the object of the verb, e.g. in a sentence such as “The control of uvrB was found to *result* from direct repression by the lexA gene product” where *the control* is the Theme of the verb *result*
	-		In passive sentences it is normally the subject of the verb
	-		It is also possible for events to have more than one Theme
LOCATION	+	in, at, on	Preposition should be included within the annotated text span;
			When there is a list of more specific entities, the first of these has been annotated only
		Near	This is a more *vague* Location; in case that a vague and a more specific locations are specified in the text, both have been annotated as two separate instances of Location semantic role
		between	Each of the entities that represent the bounding points of the location has been annotated separately as Location(s)
MANNER	+	by, through	
	-	-*ing* form	When the verb *using* precedes
	-		When the role is expressed through an adverb
	+	in	When the corresponding noun phrase ends with the word *manner* or synonym (e.g. *fashion*)
			When the preposition introduces a fixed set of latin phrases, e.g. *in vitro*, *in vivo*, *in trans*, etc.
CONDITION	+	in	As part of the phrase *in response to*, when it corresponds to changes of an environmental condition
			As part of the phrase *in the presence of* or *in the complete absence of*, when it corresponds to substances being present within the environment
	-		When the role is expressed through an adverb
DESTINATION	+	to, into	
PURPOSE	+	to	The preposition precedes the infinitive form of a verbal phrase

Semantic role	Preposition	Preposition type	Main annotation guidelines
	+	for	The preposition precedes a nominalised verb
RATE	+	by, to	
	+	in	When the preposition introduces a phrase as *in a n-fold increase*
	-		When the role is expressed through an adverb
SOURCE	+	from	
TEMPORAL	+	during, before, after, at, etc..	With prepositions that indicate time or ordering events
INSTRUMENT	+	with, by, through	
	-	*-ing* form	
DESCRIPTIVE-THEME	+	as	
DESCRIPTIVE-AGENT	+	as	

The first column in the Table above reports the semantic role types of the BELA scheme; the second column indicates whether a preposition is expected to introduce the corresponding semantic role filler; the third column shows the preposition type(s); the last column reports relevant excerpts from the actual annotation guidelines.

We exploited this list of 'prototypic' syntactic realisations of semantic arguments as another information source to guide the syntax-semantics mapping process.

5.2.3 General language repositories of semantic frames

In order to solve doubtful mapping cases, general language repositories of semantic frames containing both syntactic and semantic information were also resorted to. Amongst others, we choose to exploit VerbNet because, similarly to our case (see Deliverable 4.1, Section 4), it uses a set of frame-independent thematic roles. To be more concrete, VerbNet was used to guide the mapping process in cases like *depend*, whose extracted bio-event frame is Agent#Theme#Location# which had to be mapped onto the acquired subcategorization frame ARG1#PP-in#PP-on#. In VerbNet *depend* belongs to the class of "*rely*" verbs gathering verbs which share the same syntactic and thematic structure. Following VerbNet the Agent-Theme mapping appears to be as follows:

"Agent"<ARG1, "Theme"<PP-on.

5.2.4 Annotated corpus

The BELA corpus was taken as a further source of evidence, especially when the other information sources were not sufficient to perform the syntax-semantics mapping. The BELA corpus was particularly useful to cope with verbs that don't feature in a general language repository of frames or that may have a different syntactic realisation and different semantic properties in the biomedical domain.

For instance, the BELA corpus was resorted to in the definition of the mapping between the semantic frame Agent#Theme# and the subcategorization frame ARG1#PP-for# extracted for the verb *code*. According to VerbNet the verb *code* belongs to the class of "classify" verbs whose foreseen thematic roles are "Agent" and "Theme" and their syntactic counterpart is ARG1 (subject) and ARG2 (object). However, the BELA corpus contains sentences like *Thre recA gene could code for an antirepressor* whose bio-event annotation is reported below:

VERB: code
AGENT: Thre recA gene
THEME: for an antirepressor

On the basis of this, the Agent slot was mapped onto ARG1 and the Theme one onto PP-for.

5.3 *The mapping results*

The syntax-semantics mapping was carried out manually on the basis of the different information sources depicted above. In particular, it focussed on the 168 verbs for which both subcategorization and bio-event frames were available and resulted into 668 linked frames.

Different types of mapping were performed:

1 full mapping, where the arity of the subcategorization and bio-event frames is the same;

2 partial mapping, covering:

2.a cases in which the semantic frame contains more slots (i.e. semantic roles) than the corresponding subcategorization frame. In these cases, a mapping could be defined for a subset of the semantic roles in the bio-event frame only, e.g.

AGENT>ARG1#THEME>ARG2#LOCATION>PP-in#SOURCE>0

2.b cases in which there are subcategorized slots which do not find a semantic counterpart in the corresponding bio-event frame. This is typically the case of event frames which did not contain an explicit mention of an AGENT role which however has been reconstructed as ARG1 at the level of the subcategorization frame: this typically applies to passive sentences like *The wild-type pcnB gene was cloned into a low-copy-number plasmidin* whose syntactic representation includes a reconstructed ARG1 which doesn't correspond to any filled semantic argument of the annotated bio-event frame. In this case the mapping presents itself as follows:

0>ARG1#THEME>ARG2#DESTINATION#PP-into

2.c cases combining both types of partial mapping described above (2.a and 2.b), i.e. where the semantic frame contains more slots than the corresponding subcategorization frame on the one hand, and a reconstructed ARG1 doesn't have any counterpart at the semantic level on the other hand.

The table below summarises achieved results:

Type of mapping		Number of cases	%
Full mapping	Same arity	239	35.76
Partial mapping	2.a	123	18.41
	2.b	166	24.85
	2.c	140	20.95
Total		668	

It should be noticed that 28 extracted bio-event frames were discarded since they turned out to originate from errors during the semantic annotation process: these errors are mainly concerned with the wrong assignment of an “Agent” role instead of a “Theme” role.

In what follows the different types of performed mapping will be exemplified.

5.3.1 Full mapping

Consider the verb *modulate*, whose extracted subcategorization and bio-event frames are reported in the table below:

Verb	Extracted	
	Bio-event frames	Subcat frames
modulate	1 Agent#Theme# 2 Agent#Theme#Manner# 3 Agent#Theme#Purpose# 4 Agent#Theme#Source#	A ARG1#ARG2# B ARG1#ARG2#PP-in# C ARG1#ARG2#PP-by#

By combining the different information sources, the Agent#Theme#Manner# frame (2) was linked to the ARG1#ARG2#PP-in# (B) and ARG1#ARG2#PP-by# (C) subcategorization frames as follows:

- Agent>ARG1#Theme>ARG2#Manner>PP-by#
- Agent>ARG1#Theme>ARG2#Manner>PP-in#

It can be noticed that in both cases the prominence-preserving constraint was respected; moreover, the mapping of the Manner role was driven by the information on the prototypical realization of semantic roles in the Annotation Guidelines.

Note, however, that in some cases of full mapping the relative ordering of linked slots is not aligned. This is the case for example of a verb such as *act* whose mapped bio-event and subcategorization frames are respectively Agent#Condition#Descriptive-Agent# and ARG1#PP-as#PP-in# and whose linking result is reported below:

- Agent>ARG1#Condition>PP-in#Descriptive-Agent>PP-as#

This misalignment should not be seen as a violation to the prominence-preservation constraint since the ordering of oblique complements (typically realised as PPs) in the subcategorization frame is alphabetical.

5.3.2 Partial mapping

The following cases of partial mapping can be distinguished:

A cases in which a mapping could be defined for a subset of the semantic roles in the bio-event frame only, i.e. where the arity of the semantic frame is greater than the arity of the subcategorization frame. For example, for the verb *express*, for which the semantic frame Agent#Theme#Location#Condition# and the subcategorization frame ARG1#ARG2#PP-in# have been extracted the following mapping has been defined:

AGENT>ARG1#THEME>ARG2#LOCATION>PP-in#CONDITION>0

where it can be noticed that the semantic role CONDITION does not have an

overt syntactic realization (being equal to 0);

B the reverse of A, i.e. cases where the arity of the subcategorization frame is greater than the arity of the semantic frame; in cases like this there is one or more subcategorization slots which could not be mapped to a corresponding semantic slot. This follows from the fact that the process of subcategorization extraction has automatically reconstructed a syntactic argument ARG1 which doesn't correspond to any filled semantic argument of the annotated semantic frame. It mostly concerns verbs occurring most of the times in the passive voice and which don't have any "Agent" slot instantiated. For example, for the verb *introduce*, the following subcategorization and semantic frames have been extracted: namely, ARG1#ARG2#PP-into# and Theme#Destination#. In this case, the mapping was concerned with the "Theme"and "Destination" slots, linked to ARG2 and "PP-into" slots respectively. One important piece of evidence which was resorted to to drive the mapping process in cases like this was concerned with the percentage of times the subcategorization frame being linked was attested as occurring in the passive voice (i.e. 62% of the times). On the basis of this fact, the mapping process left unmapped the reconstructed syntactic argument ARG1;

C a third case of partial mapping is represented by a combination of both A and B above. In this case, there are both semantic slots and syntactic ones which could not find a counterpart at the other level. Consider for example the subcategorization and semantic frames extracted for the verb *delete*, respectively ARG1#ARG2#PP-from# and Theme#Source#Condition#. In this case, ARG1 on the one hand and CONDITION on the other hand do not find any linked position at the other level. As in the previous case, it appears that the subcategorization frame ARG1#ARG2#PP-from# is typically attested with the verb used in the passive voice (64% of the times).

5.3.3 Other results

It should be noted that there is a set of 219 subcategorization frames which could not be mapped to any semantic frame. There are a number of reasons for this situation to occur, i.e.

- the different size of the acquisition corpus used for subcategorization extraction with respect to the BELA annotated corpus; it should be reminded that the whole set of subcategorization frames automatically acquired have been extracted from a corpus of approximately 30,000 MEDLINE abstracts of the subject of E. Coli, while the whole set of semantic frames acquired have been manually annotated on a significantly smaller corpus of 677 of these abstracts;
- the different methods followed for the acquisition purposes, i.e. the process of subcategorization frames extraction has exploited automatic means while the annotation of semantic information has been carried out manually.

We strongly believe that this set of unmapped subcategorization frames can be used in the future to further extend the set of semantic frames associated with the selected verbs.

5.4 *Representation of syntax-semantics linking in BL*

In this section, the representation of the syntax-semantics linking in the Bio-Lexicon will be discussed, with a specific view to the interchange format to be used for uploading linking information into the Bio-Lexicon.

Event frames are encoded in the XML format reported in section 4 above. Subcategorization frames are encoded in the XML format agreed upon with WP2 team, repeated below for the reader's convenience:

```
<SubcategorizationFrame id="ARG1#ARG2#">
        <SyntacticArgument id="arg0_ARG1#ARG2#">
                <DC att="position" val="arg0"></DC>
                <DC att="function" val="subject"></DC>
                <DC att="syntacticConstituent" val="NN"></DC>
        </SyntacticArgument>
        <SyntacticArgument id="arg1_ARG1#ARG2#">
                <DC att="position" val="arg1"></DC>
                <DC att="function" val="object"></DC>
                <DC att="syntacticConstituent" val="NN"></DC>
        </SyntacticArgument>
</SubcategorizationFrame>
<SubcategorizationFrame id="ARG1#ARG2#PP-in#">
        <SyntacticArgument id="arg0_ARG1#ARG2#PP-in#">
                <DC att="position" val="arg0"></DC>
                <DC att="function" val="subject"></DC>
                <DC att="syntacticConstituent" val="NN"></DC>
        </SyntacticArgument>
        <SyntacticArgument id="arg1_ARG1#ARG2#PP-in#">
                <DC att="position" val="arg1"></DC>
                <DC att="function" val="object"></DC>
                <DC att="syntacticConstituent" val="NN"></DC>
        </SyntacticArgument>
        <SyntacticArgument id="arg2_ARG1#ARG2#PP-in#">
                <DC att="position" val="arg2"></DC>
                <DC att="function" val="comp"></DC>
                <DC att="syntacticConstituent" val="PP-in"></DC>
        </SyntacticArgument>
</SubcategorizationFrame>
```

The representation of the syntax-semantics linking was agreed with the WP2 team and is conformant to the Lexical Markup Framework (LMF) model (see Deliverable 2.1 for more details). In what follows we report an XML excerpt encoding linking information. Three different types of “SynSemCorrespondence” are reported, to

exemplify the XML encoding of the different types of mapping identified, namely full mapping (ISO_1), partial mapping of type 2a (AUG_1) and partial mapping of type 2b (RED_1).

```
<SynSemCorrespondence synsemId="ISO_1">
                <DC att="typeOf" val="ISOBivalent" />
                <SynSemArgMap  synFeature="arg0_ARG1#ARG2#"
                               semFeature="MANCU_V1EGR8_204_Agent" />
                <SynSemArgMap  synFeature="arg1_ARG1#ARG2#"
                               semFeature="MANCU_V1EGR8_204_Theme" />
</SynSemCorrespondence>
<SynSemCorrespondence synsemId="RED_1">
                <DC att="typeOf" val="ReducedTrivalent" />
                <SynSemArgMap  synFeature="arg0_ARG1#ARG2#PP-in"
                               semFeature="MANCU_V1EGR8_204_Agent" />
                <SynSemArgMap  synFeature="arg1_ARG1#ARG2#PP-in"
                               semFeature="MANCU_V1EGR8_204_Theme" />
                <SynSemArgMap  synFeature="arg2_ARG1#ARG2#PP-in"
                               semFeature="-" />
</SynSemCorrespondence>
<SynSemCorrespondence synsemId="AUG_1">
                <DC att="typeOf" val="AugmentedTrivalent" />
                <SynSemArgMap  synFeature="arg0_ARG1#ARG2#PP-at"
                               semFeature="MANCU_V1EGR81_206_Agent" />
                <SynSemArgMap synFeature="arg1_ARG1#ARG2#PP-at"
                               semFeature="MANCU_V1EGR81_206_Theme" />
                <SynSemArgMap  synFeature="-"
                               semFeature="MANCU_V1EGR81_206_Source" />
</SynSemCorrespondence>
```

Having defined the different types of “SynSemCorrespondence” holding between different extracted subcategorization and event frames for the same verb, they are then listed at the level of the lexical entry within the “Corresp” element as exmplified below:

```
<Cluster clsId="MANCU_V1EGR8" semType="MANCU_Event">
                <Entry entryId="MANCU_V1EGR8_1" baseForm="activate"
                            type="PREFERRED">
                        <PosDC posName="POS" pos="V" />
                        <PredicativeRepresentation id ="activate_1"
                                predicate="activate#agent#Theme" >
                                <Corresp ssc="ISO_1" />
...
                </Entry>
</Cluster>
```

6 Linking between BELA and BEBA corpora

UoM and EBI have collaborated to integrate the linguistic annotation and the biological annotation on a sample of the annotated corpora. The sample consists of 14 abstracts, which have been annotated with 44 biological events and with 164 linguistic events. The 44 biological events have 198 attributes (e.g. event type, participant, polarity), and the 164 linguistic events have 443 event arguments labelled with their roles in the events (e.g. Agent, Theme, Verb, Location).

We have investigated links between the two levels of annotation by linking linguistic evidence to each attribute of biological annotations. Figure 1 shows an example of biological annotation with links to the corresponding linguistic annotation, which is shown in Figure 2. Each attribute of a biological event has a tag of “<evidence ...>” in the comment field. The identifiers assigned to the tags are from the corresponding linguistic annotation. If an attribute of a biological event can be and should be linked to more than one linguistic annotations then all the identifiers of the linguistic annotations are assigned to the evidence tag of the biological event attribute. For instance, the agent of the first biological event (i.e. GcvA) is found three times in the linguistic annotation, where the mentions of the agent play the semantic role of effecting the transcriptional regulation event onto

the operon gcv. It should be noted here that *agents* at the biological level do not necessarily have to be linked to instances of the AGENT semantic role at the linguistic level. It is possible for evidence from the linguistic level annotation to correspond either to events themselves, or *any* semantic arguments associated with events.

```
<BAB>
<ROT agents="GcvA" patients="gcv" polarity="positive/negative" direct="yes"/>
<comment>
<evidence event="1,5,13"/>
<evidence agents="3,7,15"/>
<evidence patients="4,8,17"/>
<evidence polarity="1,5"/>
<evidence direct="13"/>
</comment>
<ROT agents="GcvA" patients="gcvA" polarity="negative" direct="yes"/>
<comment>
<evidence event="9,12"/>
<evidence agents="11"/>
<evidence patients="12"/>
<evidence polarity="9"/>
<evidence direct=""/>
</comment>
<SENT>The GcvA protein both activates and represses the gcv operon and negatively regulates its own transcription</SENT>
<SENT>GcvA binds to three sites in the gcv control region and to one site in the gcvA control region; each of these binding sites contains the conserved 5 bp DNA sequence 5'-CTAAT-3'.</SENT>
</BAB>
```

Table 1: Example of biological annotation with links to linguistic annotation

<SLOT argid="3" eventid="1" Role="Agent"><SLOT argid="7" eventid="5" Role="Agent"><SLOT argid="11" eventid="9" Role="Agent"><NE cat="PROTEIN">The GcvA protein</NE></SLOT></SLOT></SLOT> both <EVENT id="1"><SLOT argid="2" eventid="1" Role="Verb">activates</SLOT></EVENT> and <EVENT id="5"><SLOT argid="6" eventid="5" Role="Verb">represses</SLOT></EVENT> <SLOT argid="4" eventid="1" Role="Theme"><SLOT argid="8" eventid="5" Role="Theme"><NE cat="OPERON">the gcv operon</NE></SLOT></SLOT> and <EVENT id="9">negatively <SLOT argid="10" eventid="9" Role="Verb">regulates</SLOT></EVENT> <SLOT argid="12" eventid="9" Role="Theme">its own transcription</SLOT>.

<SLOT argid="15" eventid="13" Role="Agent">GcvA</SLOT> <EVENT id="13"><SLOT argid="14" eventid="13" Role="Verb">binds</SLOT></EVENT> to <SLOT argid="16" eventid="13" Role="Theme">three sites</SLOT> <SLOT argid="17" eventid="13" Role="Location">in the gcv control region</SLOT> and to one site in the gcvA control region; each of <SLOT argid="20" eventid="18" Role="Theme">these binding sites</SLOT> <EVENT id="18"><SLOT argid="19" eventid="18" Role="Verb">contains</SLOT></EVENT> <SLOT argid="21" eventid="18" Role="Descriptive-Theme">the conserved 5 bp DNA sequence 5'-CTAAT-3</SLOT>'.

Table 2: Linguistic annotation example

Two curators first carried out the linking task separately and then merged their annotations after discussion about the results. 118 attributes of biological events (59.6%) have been successfully linked to the corresponding attributes of linguistic annotations. This integrated corpus can serve as a training corpus for learning language patterns that link the linguistic events, which can be identified from predicate-argument structures of sentences, to the biological events, which can be directly used for database population. It is also planned to carry out further linguistic annotation to increase the overlap with EBI's biological annotation, thus providing a greater amount of training data. This work will be carried out in the context of WP11.

7 References

Baker, M.C. (1989) “Object Sharing and Projection in Serial Verb Constructions”, Linguistic Inquiry 20, 513-553.

Belletti, A. and L. Rizzi. (1988). Psych-Verbs and Theta-Theory. *Natural Language and Linguistic Theor*y 6: 291-352

Bresnan, J. and J. Kanerva (1989) “Locative Inversion in Chicheˆwa: A Case Study of Factorization in Grammar”, Linguistic Inquiry 20, 1-50.

Carrier-Duncan, J. (1985) “Linking of Thematic Roles in Derivational Word Formation”, Linguistic Inquiry 16, 1-34.

Cohen, K.B & Hunter, L. (2006). A critical review of PASBio's argument structures for biomedical verbs. *BMC Bioinformatics 7 (Suppl. 3),* S5.

Collins, P. C. (2006). Can and may: monosemy or polysemy?. In I. Mushin & M. Laughren, (Eds.), *Annual Meeting of the Australian Linguistic Society*, Brisbane, Australia.

Dixon, R. M. W. (2005) A Semantic Approach to English Grammar. Oxford: Oxford University Press.

Fillmore, Charles J. (1968): The case for case. In Bach and Harms (Ed.): Universals in Linguistic Theory. New York: Holt, Rinehart, and Winston, 1-88.

Givón, T. (1984) Syntax: A Functional-Typological Introduction I, Benjamins, Amsterdam.

Grimshaw, J. 1990. *Argument Structure*. Cambridge: MIT Press.

Hyland, K. (1996a). Talking to the Academy: Forms of Hedging in Science Research Articles. *Written Communication,* 13(2), pp.251--281.

Jackendoff, R. (1972) *Semantic Interpretation in Generative Grammar.* MIT Press, Cambridge, MA.

Jackendoff, R. (1990). Semantic Structures. The MIT Press, Cambridge, MA.

Lafferty, J., McCallum, A. and Pereira, F. (2001) "Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data", In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001),* pp 282-289.

Larson, R.K. (1988) “On the Double Object Construction”, Linguistic Inquiry 19, 335-391.

Levin, B. and M. Rappaport Hovav (1996) ”Lexical Semantics and Syntactic Structure", in S. Lappin, ed., *The Handbook of Contemporary Semantic Theory,* Blackwell, Oxford, 487-507.

Medlock, B. & Briscoe, T. (2007). Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, pp. 992--999.

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J. & Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain”. *BMC Bioinformatics* 8:50.

Speas, M. (1990). Phrase Structure and Natural Language. Dordrecht, the Netherlands: Kluwer Academic Press.

Tsai R.T.H, Chou W.C., Su Y.S., Lin Y.C., Sung C.L., Dai H.J, Yeh I.T.H., Ku W, Sung T.Y & Hsu W.L. (2007). BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features, *BMC Bioinformatics* 8:325

Van Valin, Jr.**,** Robert D. (1990). “Layered syntax in role and reference grammar”. In Layers and Levels of Representation in Language Theory, Nuyts**,** Jan, A. Machtelt Bolkestein and Co Vet (eds.), 193 ff.

Wilbur, W.J., Rzhetsky, A., Shatkay, H. (2006) New Directions in Biomedical Text Annotations: Definitions, Guidelines and Corpus Construction. *BMC Bioinformatics*. 7:356